Structure
Article

Between Order and Disorder in Protein Structures: Analysis of "Dual Personality" Fragments in Proteins

Ying Zhang,[1] Boguslaw Stec,[1] and Adam Godzik[1,*]
[1]Program in Bioinformatics and Systems Biology, Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA
*Correspondence: adam@burnham.org
DOI 10.1016/j.str.2007.07.012

SUMMARY

In their natural environment, three-dimensional structures of proteins undergo significant fluctuations and are often partially or completely disordered. This phenomenon recently became the focus of much attention, as many proteins, especially from higher organisms, were shown to contain large intrinsically disordered regions. Such disordered regions may become ordered only under very specific circumstances, if at all, and can be recognized by specific amino acid composition and sequence signatures. Here, we suggest that the balance between order and disorder is much more subtle in that many regions are very close to the order/disorder boundary. Specifically, analysis of redundant sets of experimental models of protein structures, where emphasis is put on comparison of structures of identical proteins solved in different conditions and functional states, shows hundreds of fragments captured in two states: ordered and disordered. We show that such fragments, which we call here "dual personality" (DP) fragments, have distinctive features that differentiate them from both regularly folded and intrinsically disordered fragments. We hypothesize, and show on several examples, that such fragments are often targets of regulation, either by allostery or posttranslational modifications.

INTRODUCTION

The molecular three-dimensional structure of a protein is fundamentally tied to its function; therefore, understanding protein structure plays a very important role in understanding its function. Processes such as enzymatic reactions, the recognition of substrates, and the interactions between proteins all occur on the molecular level. This is a main premise of structural biology, which has successfully provided us with molecular-level understanding of many processes in the cell. Beautifully complicated pictures of protein structures routinely appear on the covers of popular research magazines. While such images capture people's imagination, they also, erroneously, show protein structures as something solid and well defined.

However, structure is important not only in what it is, but also in how it changes (Huber, 1987). While proteins are often presented as solid, rigid bodies, they are, in reality, highly dynamic, which is often an important feature in their function and its regulation. Some fragments of some proteins, and sometimes the entire proteins, do not actually have a well-defined structure in solution, but assume such structure only in a specific functional state. Such proteins, called intrinsically disordered or unstructured, were noticed in the mid-1990s and extensively studied ever since (Dunker et al., 2001; Dyson and Wright, 2005; Tompa, 2002).

While some experimental techniques provide a direct measure of the flexibility of a protein chain, X-ray crystallography provides such information only indirectly. That is, protein fragments that are not well ordered in the crystal are simply not visible in electron density, and, subsequently, they are not built into the final model. Information on such fragments have been collected in dedicated databases and extensively analyzed for sequence features that distinguish them from the regular, ordered segments (Dunker et al., 2001). Currently, there are over 20 different algorithms for predicting disorder in proteins (Ferron et al., 2006). In blind predictions, such as those in the CASP experiment, the best groups successfully identify nearly half of the disordered residues with false positive rates less than 20% (Jin and Dunbrack, 2005).

In all previous studies of disordered segments in proteins, the analysis was conducted on nonredundant sets of PDB proteins, where one representative protein is selected to represent all its homologs and/or various experimental models. Such subsets are typically prepared by clustering the sequences of all PDB proteins at a certain similarity threshold, often 40%–60% (50% in the DisProt database [Sickmeier et al., 2007]), to avoid using related proteins in the analysis. While it is very important to avoid statistical biases in sequence analysis, removing redundancies may also remove important information such as subtle differences between models. It is exactly such differences that are the subject of this study.

In nonredundant databases, the highest quality structure, as measured by resolution and the quality of the refinement, is usually selected to represent each cluster. It

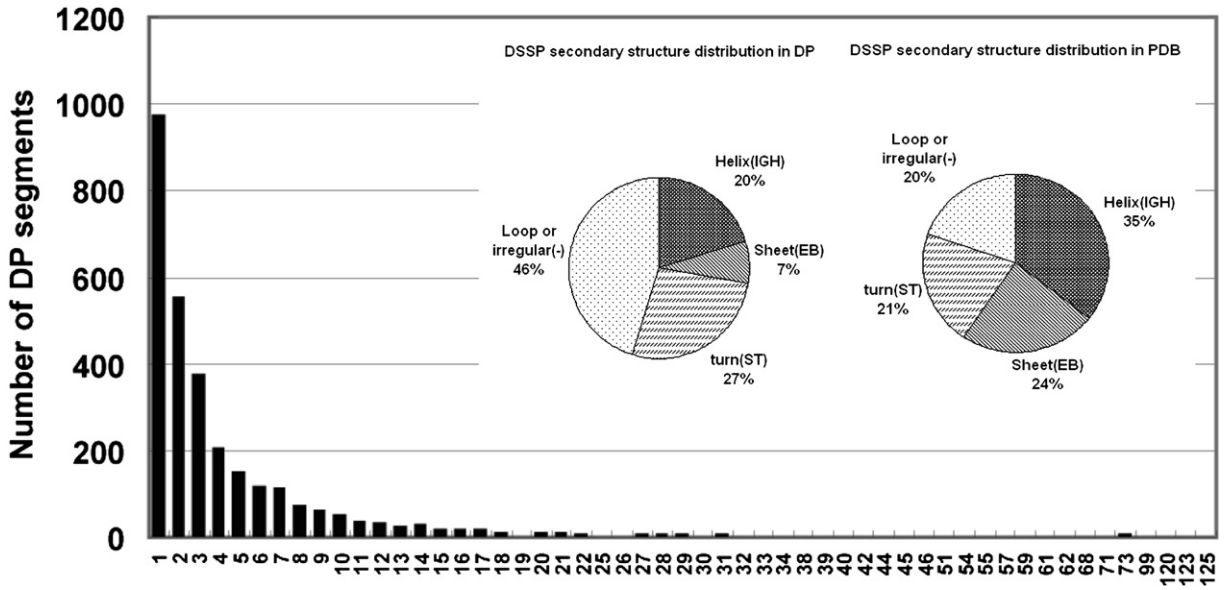

Figure 1. Length Distribution of DP Fragments

The x axis is the length of DP segments, and the y axis is the number of DP segments with given length; 92.3% of the DP segments are less than ten residues. Inset: distribution of DP fragments is different secondary structure elements as compared to the distribution in the entire database.

is customary to interpret resolution and level of refinement as qualitative factors in assessing the model. As such, models obtained from high-resolution crystals and refined to low R factors are considered to be "better." Here, we argue that this is not entirely correct. "Improving" resolution can be accomplished by introducing mutations or cofactors or by altering crystallization conditions, all of which can change the structure of a protein. Such changes, in turn, will result in better crystal packing. Therefore, differences between protein models obtained from different crystals may, at least partly, reflect actual differences between protein structures in different conditions. In this interpretation, each independently solved structure represents a slightly different experiment with subtle differences in the structure reflecting the effects of differences, such as packing, cofactor binding, or active state, on the structure of the protein in question. Therefore, studying differences between redundant structures provides us with a first glimpse into how protein structure reacts to small changes in its environment.

At the same time, some differences between independently obtained models of the same protein may reflect specific choices in data processing and in the technical procedure of model building between different research groups, and, thus, not the actual differences between structures. However, the surprisingly consistent picture that emerges from our analyses suggests that such cases, even though undoubtedly present in our database, do not influence the overall results of this study.

In this contribution, therefore, we analyze a redundant set of proteins by specifically selecting clusters of different models of identical proteins, as measured by sequence identity. We use the CD-HIT program (Li and Godzik, 2006; Li et al., 2001) to cluster all PDB proteins at 100% sequence identity. With this threshold, we disregard effects of point mutations and errors in reporting the sequence, thus probably significantly underrepresenting the scope of local structure changes. This set of proteins will be the focus of the analyses presented in this paper, as described in detail in the Experimental Procedures sections.

There are many types of differences, most of them very subtle, between "redundant" protein models. These differences include the extent of secondary structure elements, subtle shifts of individual residues, entire loops, or sometimes even secondary structure elements, and, finally, the length and positions of disordered regions (i.e., regions not present in the model at all). Such differences, occasionally mentioned in individual structure analysis papers, but to the best of our knowledge never studied systematically, are an extreme example of a much broader concept of chameleon sequences (Argos, 1987; Guo et al., 2007), where identical sequence fragments were shown to adapt different local structures (for instance, helix ↔ beta) in different protein structures (here, we study local structural differences between different models of identical proteins).

In the literature, regions not observed in protein models are associated with "intrinsically disordered" fragments. Chain fragments seen in some models, but not in others, clearly can exist either in the ordered (O) or disordered (D) states, and, therefore, we propose to call them "dual personality" (DP) fragments. In this contribution, we focus specifically on such fragments.

RESULTS

Statistics

A total of 19,858 PDB structures fulfilling the criteria outlined in the Experimental Procedures section were

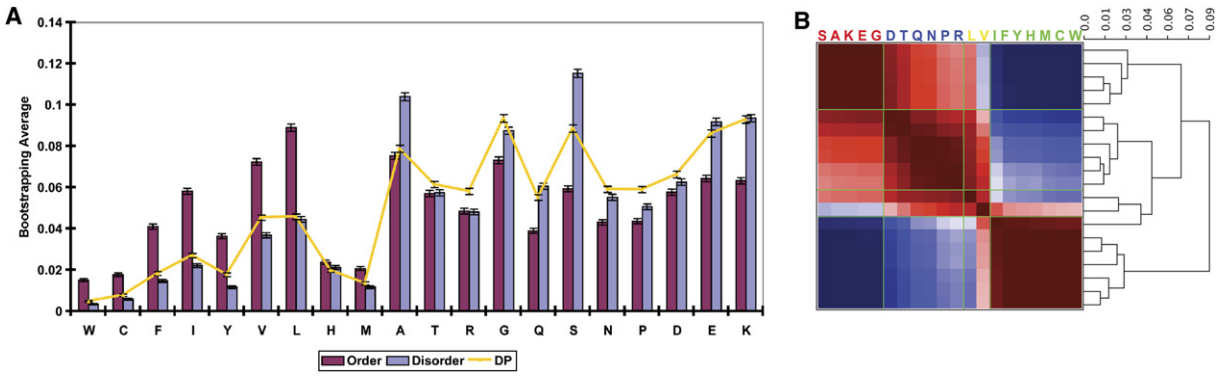

Figure 2. Comparisons of Amino Acid Frequencies of Ordered, Disordered, and DP Fragments

(A) Amino acid distribution of ordered (red), disordered (blue), and DP (yellow) fragments. Error bars are standard deviations given by the bootstrapping method.

(B) Amino acids clustered based on their frequencies in the ordered, disordered, and DP fragments.

clustered into 3,412 groups with identical sequences and more than one structure per cluster (our database also contains 2,447 clusters with only one structure per cluster) and analyzed for the presence of "dual personality" fragments. This result is an interesting comment on PDB redundancy as ~58% of all proteins deposited to the PDB have been solved at least twice in an independent experiment. More than 70% of these clusters contain disordered regions, and over 45% contain DP regions (in both cases, often several in a single protein). In total, 2,819 DP fragments were found in 1,535 clusters.

Distribution of DP Fragments

As shown in Figure 1, distribution of fragment lengths follows the power law ($y = ax^k$). About 92.3% of the DP fragments are less than ten amino acids in length. Among all the DP fragments, 50% are immediately adjacent to a disordered fragment (D), and 50% appear in an otherwise well-ordered (O) region.

Inset to Figure 1 shows the distribution of DP fragments in secondary structure elements. Intuitively, one can expect that fragments that could be disordered would be found only in loops or turns. Indeed, this is where the DP fragments are most likely to be found, but surprising 27% of DP fragments can be found in regular secondary structure elements (helix [20%] and sheets [7%]).

Amino Acid Distributions

Single Amino Acid Distribution

We use the bootstrapping method (Conrad, 1979) to estimate the amino acid distribution and its standard errors, in each of the three (O, D, and DP) classes. The standard errors give confidence intervals in comparing amino acid distributions (Figure 2A), clearly showing that amino acid distribution in DP fragments is significantly different from that found in the disordered ones, indicating that we are dealing with three different distributions and, thus, three different populations.

Specifically, in Figure 2A, the amino acids are arranged in the order of "flexibility," following the scale of Vihinen and coworkers (Vihinen et al., 1994). The tendency of amino acids to be rigid (and buried) increases to the left, while the tendency to be flexible (and exposed) increases to the right (Vihinen et al., 1994). The same arrangement is used in Figure 3.

Clustering amino acids based on the ratios between their relative abundance in ordered, disordered, and DP regions identified three major groups (see Figure 2B). Hydrophilic and small amino acids (S, A, K, E, and G) constitute the first group. They are abundant in disordered fragments and deficient in ordered fragments. From this group, A shows similar preference to ordered as to the DP fragments, while preference for S is halfway between

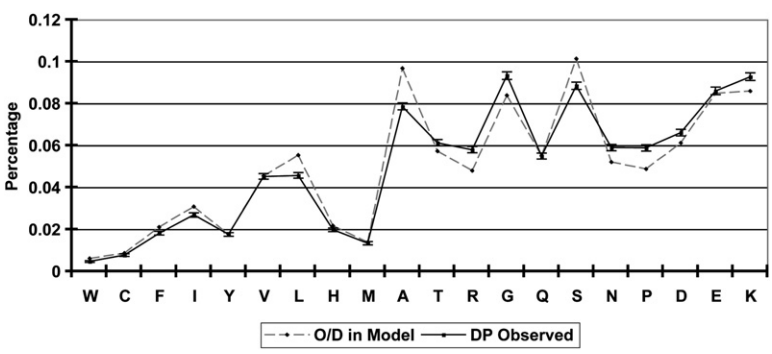

Figure 3. Amino Acid Distribution in the Hypothetical O/D State, Which Is a Linear Combination of the Ordered and Disordered States

O/D state is shown by the dashed line, with the DP state shown by the solid line. Error bars are standard deviations by the bootstrapping method.

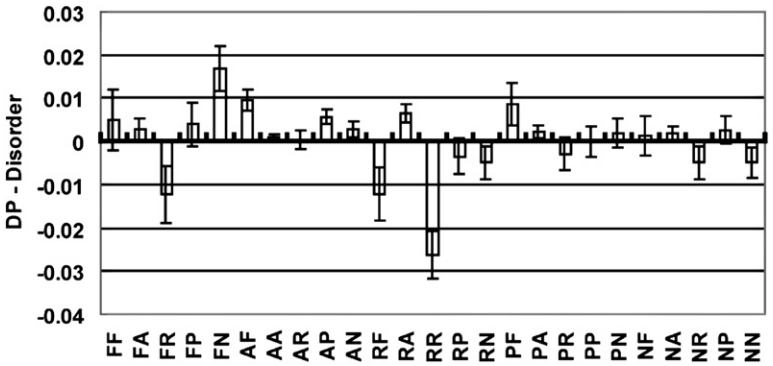

Figure 4. Analysis of Pair-Wise Distributions in DP and Disordered Fragments

A Simplified, Five-Letter Amino Acid Alphabet of Hydrophobic (F), Aromatic (A), Positively Charged (P), Negatively Charged (N), and Polar (R) is used.

The y axis shows a difference between a given pair frequency in the DP fragments minus its frequency in the disordered fragments. Error bars are calculated by the bootstrapping method.

disordered and ordered preferences. The second group of mostly hydrophilic aminoacids (D, T, Q, N, P, and R) show very similar preferences for ordered and disordered states, while most of them have higher preference for the DP fragments. The third group includes seven mostly hydrophobic amino acids (I, F, Y, H, M, C, and W). These amino acids are abundant in ordered fragments and deficient in disordered fragments. Finally, the last group (L, V) clusters mostly with the third group. Most amino acids in these two groups don't show much difference between disordered and DP preferences.

Linear Model

Features of DP fragments may be viewed as intermediate between ordered and disordered, therefore one might hypothesize that amino acid frequencies characteristic of DP fragments be a linear combination of frequencies for ordered and disordered fragments. To evaluate this hypothesis, we created a model in which a hypothetical order/disorder (O/D) state is modeled by a linear combination of ordered and disordered states. The frequency of amino acids in this state can be described by the formula:

$$\%AA_{O/D} = \lambda \times \%AA_{order} + (1 - \lambda) \times \%AA_{disorder}$$

The optimal value of λ equal to 0.25 was obtained by linear least square fitting to the observed amino acid frequencies with norm R equal to 0.03.

The comparison of DP and O/D states are shown in Figure 3. While this model correctly accounts for frequencies of about half of amino acids (mostly those very hydrophobic and very hydrophilic), the other half clearly lies outside the confidence intervals calculated by bootstrapping, especially those classified into the third group of amino acids in the last section (T, R, G, N, P, D).

Comparison of Disordered and DP Fragments

To further evaluate the unique features of DP fragments, we compared the amino acid pairwise distributions in DP and disordered fragments. To simplify the analysis, we classified the amino acids into five groups based on their physicochemical properties: hydrophobic (F), aromatic (A), positively charged (P), negatively charged (N), and polar (R). Figure 4 shows the differences of the pairwise distributions in this simplified alphabet between the DP and disordered regions. In Figure 4, the disordered region shows a clear affinity to pairs of polar residues, and the DP fragments show affinity to the pairing between hydrophobic and either positively or negatively charged residues.

Functional Features of DP

We used the ScanProsite program (Gattiker et al., 2002) to investigate further the possible functional associations of the DP fragments. Most of the DP fragments (70%) are predicted to be involved in posttranslational modification (PTM), so here, we focus on the relationship between the DP fragments and PTMs, comparing it to the relationship between PTM and both ordered and disordered fragments (Figure 5).

It is observed that DP fragments are more likely to contain PTM sites compared with ordered and disordered ones, but this effect is even more pronounced if we take into account their immediate neighborhood. To explain,

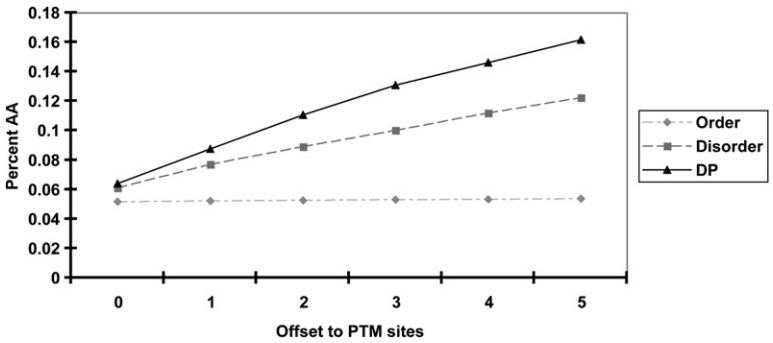

Figure 5. Analysis of Sequence Location of PTM Sites

The ordered (double dashed line), disordered (dashed line), and DP (solid line) fragments are compared. The x axis is the offset in number of residues (upstream or downstream along sequence) to the predicted PTM site. The y axis is the percentage of PTM sites taken in a specific region.

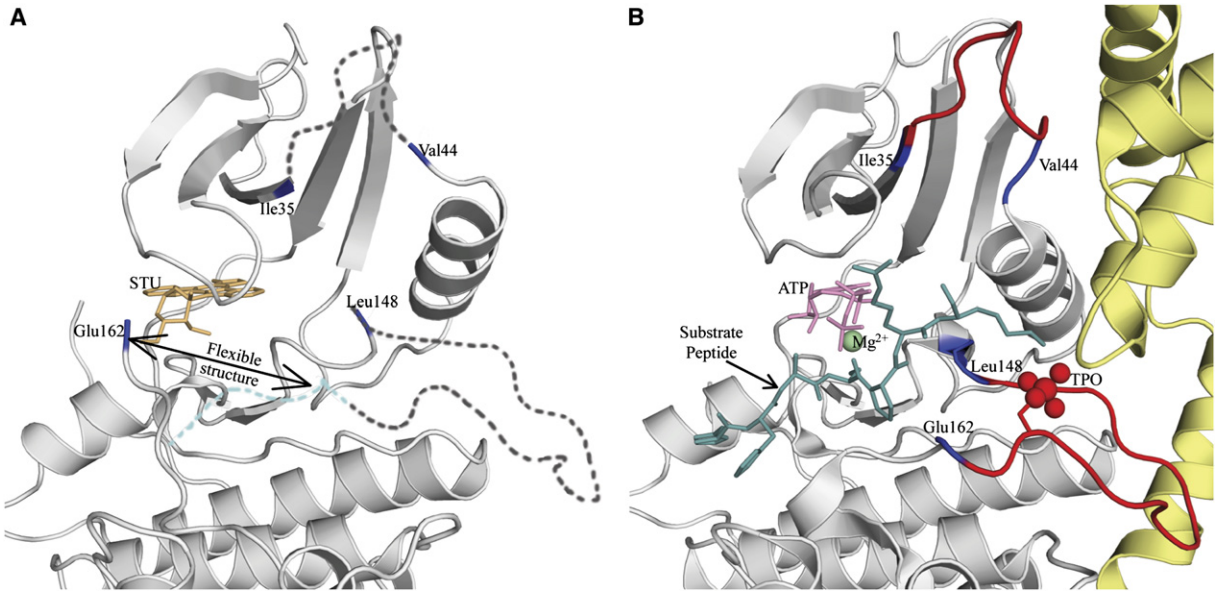

Figure 6. Comparison of the Two CDK2 X-Ray Structures

Two CDK2 structures, (A) in complex with an inhibitor, staurosporine (1AQ1:A) and (B) in complex with a substrate peptide (1QMZ:A) are being compared. Dashed curve lines denote the approximate position of the disordered fragments. STU is the inhibitor Staurosporine. TPO is the phosphorylated Threonine. The cyclin is shown in yellow on 1QMZ:A. The two DP segments are shown with dashed lines on 1AQ1:A and in red on 1QMZ:A.

predicted PTM sites are 20% more likely to be found within five residues of the DP fragment, as compared to a disordered fragment, and three times more likely than in a continuous ordered fragment (see Figure 5).

An Example of DP Fragments

In this section, we show one specific example of DP fragment identified in our database (Figure 6). Cyclin Dependent Kinase 2 (CDK2) is a well-studied protein in the protein kinase family. Over 60 structures of CDK2 solved in different conditions were deposited into the PDB (Berman et al., 2000). Figure 6 compares two representative structures (1AQ1:A and 1QMZ:A). There are two DP segments in this cluster. One segment runs from residue Ile35 to Val44; the other runs from Leu148 to Glu162. When CDK2 is solved with inhibitor staurosporine (1AQ1:A), the two segments are highly flexible and disordered (not presented in the model). However, when CDK2 is associated with a cyclin and phosphorylated by a separate protein kinase at the Thr160 residue (within the DP fragment), the two segments become ordered and are present in the model (1QMZ:A). These two effects (phopshorylation and substrate binding) are coupled with order/disorder transition in two DP fragments. Similar to the results of the previous section, this suggests that order/disorder transitions of some fragments of proteins upon signals, such as PTM, may play a role in the regulation of protein activity.

DISCUSSION

The results presented here illustrate a new aspect of the balance between order and disorder in proteins. In the currently predominant view, disorder is considered an intrinsic feature of certain relatively long protein fragments or entire proteins (Dunker et al., 2001; Dyson and Wright, 2005; Tompa, 2002). Such long, intrinsically disordered (or, as some prefer to call them, unstructured) fragments or proteins may become ordered upon, for example, dimerization, but even then, they carry specific structural signatures (Gunasekaran et al., 2004). Our analysis suggests that this division is more gradual and that there might even be a third type of fragments that literally straddles the boundary between the ordered and disordered phases, and, therefore, could be pushed one way or another by a small change in environmental conditions or protein modification, such as phosphorylation or glycosylation. Such fragments are usually short, and often, but not always, found on loops. They often influence the entire region in which they are found and represent the most visible of the conformational changes in the region. We propose to call them dual personality fragments and we argue that they can be viewed as complementing the dichotomy between the well-ordered fragments, which never change their local structures, such as fragments captured in the I-site library (Bystroff and Baker, 1998) and intrinsically disordered fragments, which are disordered in their native state and may require major changes in the environment to become ordered (Dunker et al., 2001). We can expect that other types of DP fragments, such as fragments that can change local structure from helical to extended in different environments, may also exist.

Several recent papers discuss similar concepts, such as *mo*lecular *r*ecognition *f*eatures (MoRFs) (Vacic et al., 2007), short regions in proteins that undergo order/disorder transition upon binding specific target proteins or molecules, or *e*ukaryotic *l*inear *m*otifs (ELMs), specific

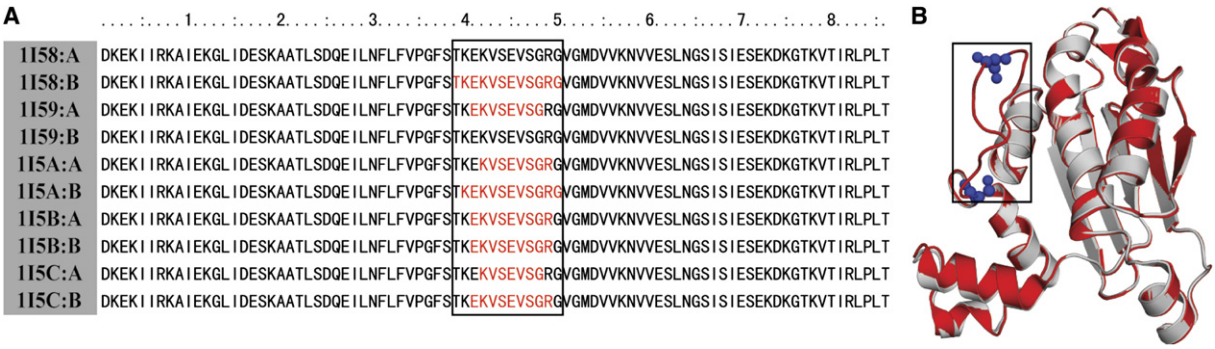

Figure 7. Visualization of DP Fragments

(A) Example of a DP fragment shown in sequence, highlighted by a box. Red, disordered; black, ordered.

(B) The same fragment, missing in the 1I58:B model (gray) but present in the 1I59:B model (red), is shown in the three-dimensional structural superposition of both models. The residues in blue are potential phosphorylation sites predicted by ScanProsite (Gattiker et al., 2002).

recognition motifs located on intrinsically disordered regions (Fuxreiter et al., 2007). Both MoRFs and ELMs focus specifically on interactions between proteins and are most likely a specific subset of the DP regions discussed in this paper. Another recent paper from the Dunker group (Le Gall et al., 2007) introduces a category of "ambiguous" regions (in contrast to ordered or disordered), which may appear to be either ordered or disordered in different protein structures and are thus very close to our DP regions. However, this not being a main focus of this paper, they applied a much simplified composition analysis and concluded "'Ambiguous' … regions … to possess the amino acid compositional biases typical of intrinsically disordered proteins" (Le Gall et al., 2007).

In contrast to the results of Dunker and colleagues, more careful comparison of amino acid distributions in ordered, disordered, and DP fragments tells us that DP fragments have their own characteristic amino acid signatures. Six amino acids, threonine, arginine, glycine, asparagine, praline, and aspartic acid, dominate DP fragments and determine their distinct conformational characteristics. Additionally, the comparison of pairwise distributions between DP and disordered fragments revealed that pairing of hydrophobic and charged residues is preferable in the DP fragments. This fact may explain why DP is more flexible (disordered) in some conditions, while more rigid (ordered) in others as such fragments would be marginally stable in both the exposed and buried parts of the structure.

In addition, the ScanProsite predictions strongly suggest that the DP fragments have a close relationship with post-translational modifications as PTMs are preferentially located at or close to DP fragments. This observation further highlights the importance of DP fragments in proteins, since regulation of protein function is increasingly becoming viewed as an integral part of cellular function.

EXPERIMENTAL PROCEDURES

Data

The database of redundant protein structures used in this study was derived from the collection of X-ray structures deposited to the Protein Data Bank (PDB) (Berman et al., 2000) on or before May 21, 2006. Different chains in the same PDB entry (labeled by the same PDB ID) are treated as separate entries.

Since very low-resolution structures are more difficult to refine, they are more likely to contain some genuine errors. Also, many of the accepted standards and quality analytic tools used in reporting X-ray data were developed only in the late 1980s; thus, structures deposited before that tend to be of very uneven quality. To minimize such effects, we have removed from the analysis all PDB entries that: (1) were deposited before 1990; (2) have resolution lower than 2.5 Å; or (3) have an R value larger then 0.25.

The database of redundant protein structures is available at (http://dper.burnham.org/); its Web interface, partly described below, provides the possibility for more detailed analysis of all the cases discussed in this paper, as well as the identification of additional examples. Also, all data files can be downloaded as a supplement to this paper.

Finding Disordered and Dual Personality Fragments

Following the approach pioneered by Dunker and colleagues, we define a protein's disordered fragment by comparing the "ATOM" and "SEQRES" records of its PDB file. The alignment of sequences parsed from "SEQRES" and "ATOM" tags was done by the "blast2-seq" program from the NCBI toolkit (Tatusova and Madden, 1999). If a residue was shown in the SEQRES record, but not in the ATOM record, we assumed this residue to be disordered (Vucetic et al., 2005) in this model. Surprisingly, more than half of the structures in the PDB contain missing coordinates, or in our nomenclature, contain disordered regions.

Some fragments are present or absent in all proteins in the cluster. Such fragments are tagged as ordered (O) or disordered (D), respectively. About 96.3% of the structure of an average protein in our database is ordered and about 2% is disordered. Fragments present in some, but not all, models in the cluster are tagged as dual personality (DP). We further note that 1.7% of the length of the average protein consists of DP fragments.

In this paper, we use all-or-nothing DP classification, meaning that if a given fragment was seen as being disordered even once it is classified as DP. We felt that practice of multiple almost identical depositions invalidates using relative frequency of ordered or disordered observations to be used as indication of the level of dual personality tendencies. It is possible that some more careful selection of redundant examples would allow us to tackle this issue better, and we intend to return to this question in the future.

Visualization of Dual Personality Fragments

A graphic interface to our database provides a simple visualization of the DP fragments, as shown in Figure 7: IDs with a gray background

on the left show the protein and chain IDs in the format "PDB_id:CHAIN_id"; sequences in one-letter alphabet are shown to the right of the protein IDs; red color identifies disordered fragments as defined above. Therefore, DP fragments are identified by the presence of black and red entries in one column—as for example, those shown in the left panel of Figure 7. The 3D structural superposition of two structures from this group (chains 1I58:B and 1I59:B) is shown in the right panel of Figure 7.

Removal of His-Tags and Other Special Cases

Many proteins are cloned and later crystallized with polyhistidine tags, a stretch of several (usually six) His residues added to the wild-type protein to facilitate purification by Immobilized Metal-ion Affinity Chromatography (IMAC) (Porath et al., 1975). Treatment of His-tags in protein models is not very consistent. As a result, they could easily be mistaken for dual personality fragments. To avoid this, the N-terminal His-containing segments were excluded from our analysis. Similarly, N-terminal Met residue had to be treated in a special way as this residue is often removed when the His-tag is added, although it is still sometimes present in SEQRES records. We have omitted all the N-terminal Met residues from our analysis. Finally, the Sel_Met residues used for structure phasing need to be treated separately due to inconsistent reporting of such residues.

Clustering of Amino Acids

To distinguish the amino acids by their preference to different fragments, we clustered the amino acids based on their distributions on DP, disordered, and ordered fragments. We used the GAP program (Chen, 2002; Wu et al., 2006) for clustering. The input data is the single amino acids distributions on the three categories of fragments: AA_{order}, $AA_{disorder}$, and AA_{DP}, respectively, and the pair-wised score of each two amino acids was the Euclidean Distances of their three-dimensional vectors defined by the inputs.

ACKNOWLEDGMENTS

The research described in this paper was supported by the P20 GM076221 grant, Joint Center for Molecular Modeling (Y.Z. and A.G.), and the National Institutes of Health grant GM64481 (B.S.). We want to thank Drs. Yuzhen Ye, Zhanwen Li, Dana Weekes, and Sri Krishna Subramanian for help in structure analysis, visualization, and development of the database.

Received: March 27, 2007
Revised: June 21, 2007
Accepted: July 4, 2007
Published: September 11, 2007

REFERENCES

Argos, P. (1987). Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. J. Mol. Biol. *197*, 331–348.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bystroff, C., and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. J. Mol. Biol. *281*, 565–577.

Chen, C.H. (2002). Generalized association plots: information visualization via iteratively generated correlation matrics. Statist. Sinica *12*, 7–29.

Conrad, M. (1979). Bootstrapping on the adaptive landscape. Biosystems *11*, 167–182.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. (2001). Intrinsically disordered protein. J. Mol. Graph. Model. *19*, 26–59.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Ferron, F., Longhi, S., Canard, B., and Karlin, D. (2006). A practical overview of protein disorder prediction methods. Proteins *65*, 1–14.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950–956.

Gattiker, A., Gasteiger, E., and Bairoch, A. (2002). ScanProsite: a reference implementation of a PROSITE scanning tool. Appl. Bioinformatics *1*, 107–108.

Gunasekaran, K., Tsai, C.J., and Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J. Mol. Biol. *341*, 1327–1341.

Guo, J.T., Jaromczyk, J.W., and Xu, Y. (2007). Analysis of chameleon sequences and their implications in biological processes. Proteins *67*, 548–558.

Huber, R. (1987). Flexibility and rigidity, requirements for the function of proteins and protein pigment complexes. Eleventh Keilin memorial lecture. Biochem. Soc. Trans. *15*, 1009–1020.

Jin, Y., and Dunbrack, R.L., Jr. (2005). Assessment of disorder predictions in CASP6. Proteins *61* (*Suppl 7*), 167–175.

Le Gall, T., Romero, P.R., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder in the Protein Data Bank. J. Biomol. Struct. Dyn. *24*, 325–342.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics *17*, 282–283.

Porath, J., Carlsson, J., Olsson, I., and Belfrage, G. (1975). Metal chelate affinity chromatography, a new approach to protein fractionation. Nature *258*, 598–599.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., et al. (2007). DisProt: the database of disordered proteins. Nucleic Acids Res. *35*, D786–D793.

Tatusova, T.A., and Madden, T.L. (1999). BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. *174*, 247–250.

Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem. Sci. *27*, 527–533.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. J. Proteome Res. *6*, 2351–2366.

Vihinen, M., Torkkila, E., and Riikonen, P. (1994). Accuracy of protein flexibility predictions. Proteins *19*, 141–149.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., et al. (2005). DisProt: a database of protein disorder. Bioinformatics *21*, 137–140.

Wu, H.M., Tien, Y. J., Chen, C. H. (2006). GAP: a graphical environment for matrix visualization and information mining. Technical Report 2006-11, Institute of Statistical Science, Academia, Taiwan.

